Intended as an article for Molecular Biology & Evolution, Discoveries Section

The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades

Pille Hallast*[1], Chiara Batini*[1], Daniel Zadik*[1], Pierpaolo Maisano Delser[1], Jon H. Wetton[1], Eduardo Arroyo-Pardo[2], Gianpiero L. Cavalleri[3], Peter de Knijff[4], Giovanni Destro Bisol[5,6], Berit Myhre Dupuy[7], Heidi A. Eriksen[8,9], Lynn B. Jorde[10], Turi E. King[1], Maarten H. Larmuseau[11,12,13], Adolfo López de Munain[14], Ana M. López-Parra[2], Aphrodite Loutradis[15], Jelena Milasin[16], Andrea Novelletto[17], Horolma Pamjav[18], Antti Sajantila[19,20], Werner Schempp[21], Matt Sears[1], Aslıhan Tolun[22], Chris Tyler-Smith[23], Anneleen Van Geystelen[24], Scott Watkins[10], Bruce Winney[25], Mark A. Jobling[†1]

* These authors contributed equally to this work.

[1] Department of Genetics, University of Leicester, UK

[2] Laboratory of Forensic and Population Genetics, Department of Toxicology and Health Legislation, Faculty of Medicine, Complutense University, Madrid, Spain

[3] Molecular and Cellular Therapeutics, The Royal College of Surgeons in Ireland, Dublin, Ireland

[4] Department of Human Genetics, Leiden University Medical Centre, The Netherlands

[5] Istituto Italiano di Antropologia, Rome, Italy

[6] Department of Environmental Biology, Sapienza University of Rome, Italy

[7] Norwegian Institute of Public Health, Division of Forensic Sciences, Oslo, Norway

[8] Centre of Arctic Medicine, Thule Institute, University of Oulu, Finland

[9] Utsjoki Health Care Centre, Utsjoki, Finland

[10] Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, USA

[11] KU Leuven, Laboratory of Forensic Genetics and Molecular Archaeology, Leuven, Belgium

[12] KU Leuven, Department of Imaging & Pathology, Biomedical Forensic Sciences, Leuven, Belgium

[13] KU Leuven, Laboratory of Biodiversity and Evolutionary Genomics, Department of Biology, Leuven, Belgium

[14] Department of Neurosciences, University of the Basque Country, San Sebastián, Spain

[15] National Center for Thalassemias, Athens, Greece

[16] Institute of Human Genetics, School of Dental Medicine, University of Belgrade, Serbia

[17] Department of Biology, Tor Vergata University, Rome, Italy.

[18] Network of Forensic Science Institutes, Institute of Forensic Medicine, Budapest, Hungary

[19] Department of Forensic Medicine, Hjelt Institute, University of Helsinki, Finland

[20] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, Texas, USA

[21] Institute of Human Genetics, University of Freiburg, Germany

[22] Department of Molecular Biology and Genetics, Boğaziçi University, Istanbul, Turkey

[23] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[24] KU Leuven, Laboratory of Socioecology and Social Evolution, Department of Biology, Leuven, Belgium

[25] Department of Oncology, University of Oxford, UK

* These authors contributed equally to this work.

† Corresponding author:

Prof Mark A. Jobling, Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Tel.: +44 (0)116 252 3427 Fax: +44 (0)116 252 3378

Email: maj4@le.ac.uk

Keywords: Y chromosome phylogeny, single nucleotide polymorphisms, targeted resequencing, Y-STRs, purifying selection

Short title: Y chromosome phylogeny

Abstract

Many studies of human populations have used the male-specific region of the Y chromosome (MSY) as a marker, but MSY sequence variants have traditionally been subject to ascertainment bias. Also, dating of haplogroups has relied on Y-specific short tandem repeats (STRs), involving problems of mutation rate choice, and possible long-term mutation saturation. Next-generation sequencing can ascertain single nucleotide polymorphisms (SNPs) in an unbiased way, leading to phylogenies in which branch-lengths are proportional to time, and allowing the times-to-most-recent-common-ancestor (TMRCAs) of nodes to be estimated directly. Here we describe the sequencing of 3.7 Mb of MSY in each of 448 human males at a mean coverage of 51 ×, yielding 13,261 high-confidence SNPs, 65.9% of which are previously unreported. The resulting phylogeny covers the majority of the known clades, provides date estimates of nodes, and constitutes a robust evolutionary framework for analysing the history of other classes of mutation. Different clades within the tree show subtle but significant differences in branch lengths to the root. We also apply a set of 23 Y-STRs to the same samples, allowing SNP- and STR-based diversity and TMRCA estimates to be systematically compared. Ongoing purifying selection is suggested by our analysis of the phylogenetic distribution of non-synonymous variants in 15 MSY single-copy genes.

Introduction

The male-specific region of the Y chromosome (MSY) has been widely exploited in studies of human evolution and population history (Jobling and Tyler-Smith 2003), but has suffered from ascertainment bias in the sequence variants studied. Also, while phylogenies constructed from such variants over the last two decades (Hammer 1995; Underhill et al. 2000; Y Chromosome Consortium 2002; Karafet et al. 2008) have been useful in defining haplogroups whose distributions can be investigated in different populations, nodes could not be dated directly from sequence variation. Consequently, dating has generally relied on use of another class of marker, Y-specific short tandem repeats (STRs). While being comparatively free from ascertainment bias because they are variable in all populations, these markers are affected by uncertainty over the appropriate choice of mutation rate (Zhivotovsky et al. 2004). There are also possible problems due to mutation saturation over long time-scales (Busby et al. 2012; Wei et al. 2013b), and to differences among STRs in repeat motif and array length, and array inhomogeneity (Ballantyne et al. 2010).

Application of next-generation sequencing (NGS) to large segments of the MSY can provide unbiased ascertainment of single nucleotide polymorphisms (SNPs) and allows detailed phylogenies to be constructed in which branch-lengths are proportional to time, allowing direct assessment of the times-to-most-recent-common-ancestor (TMRCAs) of nodes. Recently, five NGS-based trees (1000 Genomes Project Consortium et al. 2010; Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013a; Scozzari et al. 2014) have been described (Table 1), providing insights into events in human evolution and population relationships from a patrilineal perspective. However, these trees vary greatly in their sample sizes (from 36 to 1208 Y chromosomes), their number of population samples (from 1 to 9) and their representation of known lineages. Sequencing methodologies have also been heterogeneous, with consequent variation in the amount of DNA sequenced (from 1.5 to ~10 Mb) and mean coverage (from 2 × to 50 ×). In low-coverage approaches, imputation methods have been employed to infer the allelic states of missing genotypes based on the phylogeny itself, and singletons (variants present only once in the dataset) that define the lengths of

terminal branches have been poorly ascertained, with consequent likely underestimation of branch lengths (Francalacci et al. 2013).

No single study has so far combined high-coverage sequencing of multi-megabase segments of the MSY in a wide range of samples that covers the majority of the known clades of the phylogeny. Here we accomplish this, combining a phylogeny of 334 Y chromosomes (Batini et al. submitted) based on 17 populations of Europe and the Middle East, which we use elsewhere as a tool to interrogate the demographic history of European populations, with an additional 114 MSY sequences that together ensure that major clades and deep-rooting nodes are represented.

The resulting phylogeny, based on a mean coverage of 51 × in 3.7 Mb from each of 448 Y chromosomes, contains 13,261 high-confidence SNPs. It resolves polytomies, provides date estimates for deep nodes, and constitutes a robust evolutionary framework for analysing the history of other classes of mutation affecting the MSY, including Y-Y and X-Y gene conversion events and structural variants. We also analyse variation at a set of 23 Y-STRs in all 448 samples, allowing a systematic comparison of SNP- and STR-based diversity and TMRCA estimates. Analysis of damaging non-synonymous variants in 15 single-copy genes with our sequenced regions shows an under-representation of shared variants, implying that purifying selection is active on MSY.

Results

Elsewhere (Batini et al. submitted) we have described a NGS-based MSY phylogeny based on 5996 SNPs ascertained in 334 human Y chromosomes comprising 17 population samples from Europe and the Near East, focused on illuminating the origins and histories of European patrilineages. Here, we supplemented those data with additional MSY sequences from random selections of 20 males from each of two HapMap populations, YRI (Yorubans from Ibadan, Nigeria) and CHB (Han Chinese from Beijing), plus 74 males from various populations, known from previous analyses to carry Y chromosomes belonging to specific haplogroups, in order to ensure that major clades and deep-rooting nodes were represented. Table S1 lists all the samples analysed and their populations of origin. We simultaneously generated orthologous MSY sequences from 22 great ape males using the same experimental approach, which we use here as an outgroup to the human sequences, and which will be described fully elsewhere.

We used a sequence-capture design (see Materials & Methods) that allowed us to analyse 3.7 Mb of readily interpretable human MSY sequence, excluding the ampliconic and X-transposed regions (Skaletsky et al. 2003) of the chromosome (Figure 1; Tables S2, S3), and gapped due to the repeat-masking required when designing sequence-capture probes. Mean coverage was 51 ×, and we called all SNPs (ignoring indels) with ≥6 × coverage, validating *in silico* by comparison with published whole-genome sequence and genome-wide SNP chip data. The high coverage and high threshold for variant calling led respectively to low false-negative and false-positive rates (Supplementary Material). We ascertained 13,261 SNPs, cross-referencing them with those identified in other published studies (Figure 2; Table S4): of our SNPs, 2356 (17.8%) are in dbSNP build 138, 8742 (65.9%) have not been previously reported, and over half are singletons, i.e. unique in the dataset (7782; 58.7%).

Features of the MSY phylogeny, and TMRCAs of nodes

A maximum-parsimony tree (Figure 3; Figure S1) was built using the program PHYLIP (Felsenstein 2005), and rooted using great ape sequences generated using the same

technical approach as the human sequences (see Supplementary Material). We used the program AMY-tree (Van Geystelen et al. 2013a) to seek previously-identified haplogroup-defining SNPs within our dataset, and thus to name the major clades based on existing Y-haplogroup nomenclature (Karafet et al. 2008); for all 154 samples for which previously generated Y-SNP data were available, all such designations were consistent. We also applied a new version of AMY-tree (version 2.0) that considered only the targeted MSY regions to assess phylogenetic consistency, and thus to act as an additional quality-control for our data (Supplementary Material). This approach revealed no calls ascribable to sequencing error in a total of 516,096 genotypes at 1152 sites checked, confirming the high data quality. Our phylogeny contains all known top-level alphabetically-labelled clades, with the exceptions of haplogroups A00 (Mendez et al. 2013), M and S. We used the rho statistic (Forster et al. 1996; Saillard et al. 2000) to estimate TMRCA for major nodes within the tree (Table 2). Previously we (Batini et al. submitted) and others (Wei et al. 2013b; Scozzari et al. 2014) have also employed the coalescent-based method implemented in BEAST (Drummond et al. 2005; Drummond and Rambaut 2007), but here our sampling violates the requirement of population sampling, so we focus on rho, noting that dates for NGS data estimated using both methods are strongly correlated (Scozzari et al. 2014; Batini et al. submitted). We use a pedigree-based MSY-specific mutation rate (Xue et al. 2009), and address the issue of mutation rate choice in the Discussion.

Here we comment on striking or novel features of the phylogeny and the point estimates of TMRCAs, beginning with the deeper-rooting nodes and then focusing on some specific clades:

(i) ***Basal African clades:*** The tree demonstrates the remarkable depth of MSY ancestry retained among hunter-gatherer groups in Africa within the rare haplogroups A and B. As has previously been observed (Poznik et al. 2013; Scozzari et al. 2014) the longest internal branches in the entire phylogeny (Figure 3b) are those among these clades, together with the branch that separates them from the remainder of the tree, superhaplogroup DR, corresponding to the expansion of Y chromosomes following the out-of-Africa migration

(Underhill et al. 2000; Wei et al. 2013a). Considering the point estimate of mutation rate, TMRCA for the entire tree is ~126 KYA, and that for the DR node 49 KYA, which correlates reasonably well with the date of the colonisation of Eurasia.

(ii) ***Ancient population expansion:*** Within clade DR of the tree lies a deep Paleolithic lineage radiation, giving rise to haplogroups G, HF5, IJ, LT, NO and P, dating to between 23 and 33 KYA (we note that a single variant identified elsewhere (Poznik et al. 2013) resolves the polytomy of haplogroups G and H, with G branching earlier).

(iii) ***Ancient subclades within hgs C and D:*** The phylogeny also contains sequences within the largely Asian haplogroups C and D, which have not been sequenced elsewhere. In both cases, most branches are long, and the TMRCAs for the clades are similar, at 39 and 34 KYA.

(iv) ***Bantu-speaking populations and expansions in hg E:*** Within haplogroup E the most striking features are the shallow star-like genealogies within E1b1a, which predominate in the food-producing, Bantu-speaking YRI+LWK, and present a stark contrast to the ancient hunter-gatherer lineages in A and B. Hg E1b1a (here given a TMRCA of 6.9 KYA) has previously been associated with the expansion of Bantu languages, which spread widely from Central Africa ~3 KYA together with farming and iron-working (Berniell-Lee et al. 2009).

(v) ***A novel hg F sublineage associated with hg H***: One Nepalese sample had been previously assigned to hg F*, and here its branch (newly named F5) arises, with hg H, from a deep-rooting bifurcation with TMRCA 32 KYA.

(vi) ***Hg H in Asia and English Romany:*** The tree contains six sequences within haplogroup H, with a TMRCA of 27 KYA. The clade is largely found in the Indian sub-continent, but is also typical of European Roma, who originated in a founder population from north/northwestern India ~1.5 KYA (Mendizabal et al. 2011). One MSY sequence belongs to an English Romany male, previously assigned to haplogroup H1a (TEK, unpublished data), and here arising from a trifurcation with Turkish and Nepalese haplotypes.

(vii) ***Star-like expansion within hg I***: Haplogroups I and J divide at 31 KYA, and each then divides in two at similar times of 21 – 23 KYA. Within hg I1 is a striking star-like genealogy dating to 3.5 KYA (Batini et al. submitted).

(viii) ***Rare deep-rooting hg Q lineages in NW Europe***: Hg Q has been most widely investigated in terms of the peopling of the Americas from NE Asia (Karafet et al. 1999). Here, as well as an example of the common native American Q-M3 lineage, we included examples of rare European hg Q chromosomes. One of the English chromosomes belongs to the deepest-rooting lineage within Q (Q-M378) and may reflect the Jewish diaspora (Hammer et al. 2009); the other is distantly related, shares a deep node with the Mexican Q-M3 chromosome, and has an STR-haplotype closely related to those of scarce Scandinavian hg Q chromosomes (unpublished data).

(ix) ***Structure within the west Eurasian hg R***: The TMRCA of hg R is 19 KYA, and within it both hgs R1a and R1b comprise young, star-like expansions discussed extensively elsewhere (Batini et al. submitted). The addition of Central Asian chromosomes here contributes a sequence to the deepest subclade of R1b-M269, while another, in a Bhutanese individual, forms an outgroup almost as old as the R1a/R1b split.

SNP-based discrimination among males, and comparison with Y-STRs

As with other NGS studies of MSY, our analysis reveals very high diversity compared to previously established phylogenies (Karafet et al. 2008). However, despite this high resolution, not all Y chromosomes in the sample can be distinguished. The 448 MSY sequences belong to 440 different SNP haplotypes, identical cases being found in eight pairs of individuals. One pair, within hg A1, belongs to a previously reported deep-rooting English pedigree (King et al. 2007), and the males are separated by just 13 generations. The remaining seven pairs are apparently unrelated, but each pair belongs to a single population – there are three identical Saami pairs (two within hg N1c1 and one in hg I1), two Palestinian pairs (in hgs E1b1b and G2a), one Bakola pair (hg A1b), and one Italian pair (hg R1b).

The traditional tools for discriminating between closely related Y chromosomes are Y-STRs. Identification of rapidly-mutating STRs (RM-STRs) (Ballantyne et al. 2010) discriminates between brothers in 60% of cases (Ballantyne et al. 2012), so their application is expected to exceed the feasible resolution of NGS approaches. However, traditionally

applied sets of STRs have lower average mutation rates than these. To investigate the relative discrimination power of SNP and STRs, we typed all 448 samples using PowerPlex Y23 (Promega), a 23-STR forensic multiplex kit that contains markers with varying mutation rates, including two RM-YSTRs (DYS570 and DYS576)(Purps et al. 2014). The resulting STR haplotypes also fail to discriminate among all samples, yielding six identical pairs (Figure S2a), again each within-population (Table S5). Only two of these are also SNP-identical, the others being discriminated by 1 - 5 SNPs. Removal of the two RM-YSTRs leads to identical haplotypes in two additional pairs and a trio, separated by 1-3 SNPs, and still within-population. Removal of a further four STRs from the haplotype to reduce it to the 17 STRs contained in the Yfiler® kit (Life Technologies) leads to identical haplotypes in an additional trio and three pairs of individuals, including two cross-population cases (Norway-Serbia, and Serbia-Spain), and up to 31 SNPs separating members of a pair. This emphasises the homoplasic nature of STR haplotypes (Larmuseau et al. 2014), and the importance of analysing many STRs for forensic identification and genealogical purposes.

Comparison of SNP- and STR-based TMRCA estimates

Y-STRs have also been widely employed for dating purposes, and here we used our Y-STR data to estimate TMRCAs (Table S6) for the lineages dated with SNPs, allowing us to compare the two marker types. We explored three variables:

(i) Two different dating methods - rho (Forster et al. 1996; Saillard et al. 2000) and average-squared distance (ASD) (Goldstein et al. 1995a; Goldstein et al. 1995b). Each was used with either an ‘ancestral haplotype’ or ‘modal haplotype’ as a root.

(ii) Three different sets of STRs – the maximum usable set of 21 STRs (excluding only the two copies of DYS385), the same set minus the RM-YSTRs and two other loci (DYS389II, DYS448) that are potentially problematic because of complex or interrupted repeat array structure (total of 17 STRs), and finally a minimal set of 13 STRs representing the Yfiler® set minus DYS385ab, DYS389II and DYS448.

(iii) Two different mean STR mutation rates: a slow 'evolutionary' rate based on population comparisons (6.9 x 10^{-4} / STR / generation (Zhivotovsky et al. 2004)), and a faster 'pedigree' rate based on direct detection of mutations in father-son pairs (depending on the subset of STRs, 2.797 – 4.238 x 10^{-3} / STR / generation; www.yhrd.org).

Table S7 summarises the results of comparing SNP- and STR-based TMRCA estimates for a range of nodes: generally, the STRs perform poorly, giving a wide variety of TMRCAs for nodes with similar SNP-based dates, and correlation coefficients consistently below 0.6. Considering the variables described above: (i) ASD generally outperforms rho, and choice of rooting method (ancestral or modal) makes little difference. For rho, rooting via the ancestral haplotype performs much worse than via the modal haplotype; (ii) removal of RM-YSTRs, and STRs showing repeat array complexity, does not have a major influence on relationships between SNP- and STR-based estimates of TMRCA, and the effects depend upon how the root is specified; (iii) The evolutionary STR mutation rate consistently overestimates, and the pedigree rate underestimates, the TMRCAs of nodes (Figure 4a). As expected, the pedigree mutation rate performs better for young nodes (<10 KYA; Table S6), while the evolutionary rate performs better for older nodes.

SNP recurrence and branch length heterogeneity

Of the 13,261 SNPs, 123 (0.92%) are recurrent within the tree. This is significantly lower ($p<0.0001$, Chi square with Yates correction) than one recent study (172/5865; 2.9% (Wei et al. 2013a)), but significantly higher ($p<0.0002$) than another (4/2386; 0.2% (Scozzari et al. 2014)). Setting aside disparities in the number of individuals sequenced, these differences seem likely to be due to the sequencing strategies and the regions of the Y chromosome analysed in the different studies. The first (Wei et al. 2013a) is based on a whole-genome sequence dataset and therefore includes repetitive elements masked in our study in which mapping may be problematic, potentially increasing the number of apparently recurrent variants. The second (Scozzari et al. 2014) is expected to reduce such mapping problems because it employs repeat-masking, as well as excluding much XY-homologous

material covered in our study, in which recurrent gene conversion from the X chromosome (Rosser et al. 2009; Trombetta et al. 2010) may be active. A total of 294 events occur at the 123 recurrent sites within our tree. Of these events, 66 (22%) occur at CpG dinucleotides, where base-substitution rates are enhanced due to cytosine methylation. The remainder may include examples of XY gene conversion, and will be investigated elsewhere.

Visual inspection of the tree (Figure 3) shows apparent heterogeneity in branch lengths between (and also within) clades – for example, the tips of hg C sequences appear to extend further than those of other haplogroups. Also, one previous study has shown a significantly reduced mean number of mutations to the root for haplogroup A, compared to other lineages (Scozzari et al. 2014). One possible trivial cause of variation is tissue source for the DNAs – for example, our samples include lymphoblastoid cell-lines (LCLs) in which some somatic mutations might be expected to have accumulated (Wei et al. 2013a), in addition to blood and buccal samples. A comparison of the mean number of mutations to the root of the tree for the three different tissue sources (Figure S3) shows that MSY sequences in the LCLs analysed here indeed carry significantly more mutations (mean of 471, n=152; p=0.00124, one-way ANOVA) than the sequences from blood (mean of 468, n=208) or saliva (mean of 466, n=88). If somatic mutations are contributing to the branch lengths for LCL samples, we would expect these mutations to be found exclusively among the singleton mutations in terminal branches. The star-cluster of 44 MSY sequences found within hg R1b provides a means to test this, and given that it is comprised of 23 LCL and 21 non-LCL samples, has 87.5% power to detect a difference of 3 mutations in branch length. However, a comparison between the two sample types (Table S8) finds no significant difference (p=0.73; Mann-Whitney U test). This apparent discrepancy may be explained by differences among the 152 analysed LCLs in the number of passages since the cultures were first established. Considering branch lengths to the root, absolute differences between sample sources are small, so have a minimal effect on TMRCA estimates.

To address the possibility of haplogroup-specific effects, we compared the mean number of mutations to the root of our tree for 17 different major haplogroups (Table S8).

Numbers of samples per haplogroup vary widely, and once this is taken into account only two comparisons, hg E and hg O vs. hg R1b, retain any signal of distinctive branch lengths – for hg E, 56% of p-values associated with Mann-Whitney U tests on sub-sampled sequences (see Materials & Methods) were significant, and for hg O the value was 87%. To ask if this could be explained by the tissue-source effect described above, we repeated the comparisons within either LCL or non-LCL sources for these three haplogroups (Table S8). In fact, the haplogroup-specific signal is strengthened – for hg E in LCLs, 97% of p-values associated with Mann-Whitney U tests on sub-sampled sequences were significant, and for hg O the value was 100%. For non-LCL samples, the proportion of significant p-values in each case is 100%. We therefore conclude that subtle haplogroup-specific effects on branch length do exist.

Putative functional variants and evidence for purifying selection

The regions sequenced here contain 15 of the 17 single-copy MSY protein-coding genes (the missing two lie within the X-transposed region, which was not covered; Figure 1); we therefore examined variation within the coding sequences of these genes.

The 13,261 variants include 80 exonic substitutions (in 13/15 genes), of which 46 are non-synonymous (Table S9). To assess the possible effect of natural selection on these 46 variants, we used SIFT and PolyPhen2 to predict those that were damaging to protein function, and then asked whether the proportion of singletons among damaging variants was over-represented compared to the proportion (7782/13,261) in the dataset as a whole. For SIFT predictions, the difference is significant (17 variants, 16 of which are singletons; p=0.0065). For PolyPhen2 predictions, the difference is marginally non-significant (15 variants, 13 of which are singletons; p=0.0527); however, notably one doubleton variant is present in two Bakola samples that are sequence-identical (Table S5), and carry STR haplotypes differing by only three mutational steps at a single STR marker. This very close relationship of the two MSY sequences indicates that they have had very little time to be

exposed to selective effects independently from each other. Taken together, these findings support the idea that purifying selection is acting on single-copy MSY genes.

Discussion

The application of next-generation sequencing is revolutionising our picture of MSY diversity. Including our study, the five most recent NGS analyses summarised in Figure 2 and Table 1 have yielded a total of 33,479 SNPs. This tsunami of MSY variants is likely to continue, as previously unexamined populations and lineages are subjected to NGS. The 1000 Genomes Project has already contributed many more variants (Rocca et al. 2012), and a major imminent additional contribution will come from the Project's analysis of ~1250 male genomes, as well as from other sequencing projects carried out for medical genetic purposes. MSY data from these projects, like that of the Sardinian population (Francalacci et al. 2013), will be at low coverage, and therefore singleton variants will be under-represented, so terminal branch lengths may be artificially short. Sequence capture has the advantage of high coverage and good singleton representation, but unlike the MSY data from whole-genome sequencing projects, does not come for free. Our sequence capture design yielded 3.7 Mb of usable sequence for phylogeny construction, but other designs (Poznik et al. 2013) yield more than twice as much, and indeed this appears to be the approach applied by commercial suppliers of genotyping services, which offer MSY resequencing for genealogically-minded clients that will lead to many citizen-scientist generated SNPs. Currently, the nomenclature system for MSY haplogroups and variants is unstable (van Oven et al. 2014), and given all this activity there is urgent need for systematic and agreed approaches to cataloguing, validating and naming MSY variants and lineages. In order to understand the time-depths and branching orders of different parts of the MSY phylogeny, better sampling of populations and lineages is required, and given the geographical bias of citizen-science participants this will likely be driven by academic research programs.

MSY mutation rate

Although the relative ages of clades in the MSY phylogeny can now be well established thanks to the large number of variants, the absolute estimates of TMRCA remain uncertain because of corresponding uncertainty about choice of the appropriate mutation rate.

Indeed recent published estimates of equivalent nodes based on NGS data vary considerably, but this is mostly ascribable to differences in assumed mutation rates. Here, we favoured a rate (1.0×10^{-9}/bp/year) estimated directly from NGS analysis of MSY sequences in a deep-rooting pedigree (Xue et al. 2009). Though the direct nature of the analysis and the proven transmission of newly arising variants are positive features, the study's major disadvantage is that its mutation rate rests on only four observations. These numbers will improve as other resequencing studies are published, but meanwhile other studies (Mendez et al. 2013; Scozzari et al. 2014) have taken the genome-wide *de novo* mutation rate (based on a larger number of observations) and scaled it to account for male-specific transmission, thus inferring slower rates of 0.62×10^{-9} (Mendez et al. 2013) or 0.64×10^{-9} (Scozzari et al. 2014). Criticism of this approach (Elhaik et al. 2014) has been based on its indirect nature, and the fact that the resulting rates are at odds with phylogenetic mutation rate estimates ($1.5 – 2.1 \times 10^{-9}$ /bp/year (Skaletsky et al. 2003; Kuroki et al. 2006)) based on human-chimpanzee MSY comparisons. Calibration based on archaeological dates and assumptions about colonisation history (such as the peopling of the Americas (Poznik et al. 2013) or of Sardinia (Francalacci et al. 2013)) has also been applied, although it introduces other sources of uncertainty. Further analysis of deep-rooting pedigrees, combined with accumulating data on well-dated ancient DNA, should help to give more reliable mutation rate estimates in the near future.

Visual inspection of the phylogeny suggests that there may be branch length heterogeneity within our phylogeny. However, after adjustment for sample size differences, statistical support for such differences remains for only two comparisons, hg O vs hg R1b, and hg E vs hg R1b. A truly haplogroup-specific effect of this kind would imply a cis-acting factor on MSY influencing mutation directly, and this seems improbable given what is known about MSY genes. A second possibility could be a factor acting over many generations in a particular geographical region or population within which a haplogroup was frequent. Such a factor could be genetic, environmental or cultural – one possibility could be higher or lower average paternal age in a particular region, which might increase or decrease effective mutation rate for locally prevalent haplogroups. If this were the case, then we might expect

haplogroups that associate together to be similarly affected: future sequencing data on larger sample sizes should allow this to be tested.

STR-based TMRCA estimation

Data presented here and elsewhere (Wei et al. 2013b) indicate that, despite their widespread use, STRs generally perform poorly in estimating the TMRCA of haplogroups. Our expectation was that removal of STRs with particularly high mutation rates or complex internal structures might improve the performance of STR sets in TMRCA estimation. However, this was not borne out, and choice of STRs appears to make little difference. Applying the widely-employed 'evolutionary' STR mutation rate (Zhivotovsky et al. 2004) leads to systematic overestimation of TMRCAs compared to SNP data (though this is no longer true for all nodes if we apply a slower SNP mutation rate (Mendez et al. 2013); Table S6); the much faster 'pedigree' STR rate leads to underestimation generally, but performs better for younger clades. This probably reflects the increasing importance of back-mutation in older clades. Despite the diminishing cost of NGS, it seems likely that researchers will wish to continue to use STRs in dating; in order to provide a rational framework, careful analysis of large datasets comprising multiple STRs and MSY sequences will be needed. The 'citizen-scientist' community, which now generates 111-locus STR haplotypes combined with ~10-Mb MSY NGS data, may be best placed to do this.

Purifying selection and MSY gene function

Our analysis of the frequency distribution of damaging variants in MSY single-copy genes suggests that purifying selection is ongoing, and that past claims of terminal degeneration of the Y chromosome are exaggerated. These findings are consistent with the picture of long-term conservation of genes from analyses of mammalian Y chromosomes (Bellott et al. 2014), as well as previous human MSY gene resequencing (Rozen et al. 2009), and general MSY sequence diversity considerations (Wilson Sayres et al. 2014). We considered only nucleotide substitutions in our analysis, and reliable indel calling is needed to

provide a more thorough analysis. While evidence is mounting that purifying selection is acting on MSY protein-coding genes, more work is required to understand their functional roles. Candidate genes are currently lacking for some established MSY-linked phenotypes such as HIV-AIDS progression (Sezgin et al. 2009) and coronary artery disease susceptibility (Charchar et al. 2012), and there is a clear need to understand the roles of non-coding RNA genes on the MSY, as well as the suite of protein-coding genes.

Molecular evolutionary applications of high-resolution MSY phylogenies

Previously, we and others have taken a phylogenetic approach to analysing the mutational history of other classes of Y-chromosomal variants, including structural rearrangements (Repping et al. 2006; Jobling et al. 2007; Balaresque et al. 2008a; Balaresque et al. 2008b), intrachromosomal gene-conversion events (Rozen et al. 2003; Bosch et al. 2004; Hallast et al. 2013; Balaresque et al. 2014), and gene-conversion between the X and Y chromosomes outside the pseudoautosomal regions (Rosser et al. 2009; Trombetta et al. 2010; Trombetta et al. 2014). Such analyses require both a reliable phylogeny and a means of assaying the complex variants. NGS can now provide phylogenies of very high resolution, in which almost all males in a sample can be distinguished, and the phylogenetic and temporal relationships between their MSY sequences can be described in a fine-grained and unbiased way. In principle, NGS can simultaneously identify the associated complex variants. However, using NGS data to unambiguously determine the allelic states of variants in highly similar regions within the MSY, and between the MSY and the X chromosome, is challenging. Overcoming these difficulties will lead to unprecedented illumination of the complex mutational history of the Y chromosome.

Materials and Methods

Samples

DNA donors were recruited with informed consent. Human DNAs (Table S1; Supplementary Material) were extracted from saliva (using the Oragene kit), LCLs, or peripheral blood. Twenty males from each of 19 populations were supplemented by 77 assorted samples chosen based on prior Y haplogroup information. Four HapMap populations (CEU, TSI, YRI, CHB) were included, both as part of the population dataset, and to provide data on externally analysed samples for validation purposes. Two non-HapMap individuals were subsequently identified as females and removed from all downstream analysis, reducing the final number of sequenced males to 455.

Sequencing, data analysis, variant calling and filtering

For details of all procedures, see Supplementary Material.

Briefly, 3-5 μg of genomic DNA was used for library preparation and target enrichment, followed by paired-end 100-bp Illumina sequencing. Reads were mapped to the human genome reference (GRCh37), followed by local realignment, duplicate read marking and base quality score recalibration.

Variant calling and filtering was carried out leading to a final analysed region of 3,724,156 bp. In total 19,276 raw variants were called from 455 samples, and following filtering 13,261 sites in 448 samples were retained for all downstream analyses. *In silico* validation was done using Complete Genomics whole-genome sequence data (8 samples) and Omni2.5 BeadChip genotype data (88 samples). Based on the Complete Genomics comparison, the false positive error rate was 0% and false negative error rate 0.009%; more details, including genotype-based error rates, are given in Supplementary Material and Table S10. All variants have been submitted to dbSNP (Table S4), and a vcf is available from (https://www2.le.ac.uk/departments/genetics/people/jobling/publications).

Phylogenetic inference and dating

Maximum parsimony trees were created in PHYLIP v3.69 (Felsenstein 2005) and visualised using FigTree v1.4.0 (Rambaut 2006-2012). Ancestral states were defined using information from the phylogeny and from sequence data from 22 male great apes, generated concurrently with the human sequencing (Supplementary Material). For assignment of variants to branches, see Table S11 and Figure S4.

TMRCA and its standard deviation was estimated for clades within the PHYLIP outfile using the rho statistic (Forster et al. 1996; Saillard et al. 2000) implemented in a Perl script. A scaled rate of one mutation per 268.5 years was used, based on 1.0×10^{-9} mutations/nucleotide/year (Xue et al. 2009) and the number of nucleotides in our regions of interest (3,724,156). We assumed a generation time of 30 years. In addition, to capture the uncertainty in the published mutation rate we calculated TMRCA based on the bounds of its 95% confidence interval: $3.0 \times 10^{-10} – 2.5 \times 10^{-9}$ mutations/nucleotide/year (Xue et al. 2009).

Known Y-SNPs were sought using AMY-tree v1.2 (Van Geystelen et al. 2013a; Van Geystelen et al. 2013b), and v2.0 of the same software (with the option of specifying MSY sub-regions) was used for variant validation via phylogenetic consistency.

Y-STR analysis

23 Y-STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385ab, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, GATAH4, DYS481, DYS533, DYS549, DYS570, DYS576, and DYS643) were typed in all samples using the PowerPlex® Y23 system (Promega) according to manufacturer's instructions. TMRCA was calculated using 21 STRs (omitting the bilocal DYS385ab), 17 STRs (additionally omitting the RM-YSTRs DYS570, DYS576 and the complex STRs DYS389II, DYS448), or 13 STRs (additionally omitting the non-Yfiler® loci DYS481, DYS533, DYS549, DYS643). Dating methods were rho, implemented within the program NETWORK 4.612 (Bandelt et al. 1999), and average squared distance (Goldstein et al. 1995a; Goldstein et

al. 1995b). Further details are given in Supplementary Material.

Branch length heterogeneity testing

Differences in branch length across the 17 haplogroups were assessed with a pairwise comparison using a Mann–Whitney U test. Bonferroni correction was also applied to account for multiple pairwise tests. For haplogroups with n >10, ten random individuals were sampled to account for sample size variation. This process was repeated 100 times producing 100 p-values for each comparison. The proportion of significant p-values was then calculated and only comparisons with a proportion > 50% were considered of interest.

Genes and functional variants

Variants within UCSC genes within the sequenced regions were identified and their likely functional effects analysed using wANNOVAR (Chang and Wang 2012).

Acknowledgements

We thank all DNA donors; Lorna Gregory and the Oxford Genomics Centre for library preparation, target enrichment and sequencing; Diego Forni with help preparing Figure 1; Emma Parkin and Denise Carvalho-Silva for unpublished information on Nepalese and Bhutanese males. We also thank a careful reviewer for very helpful comments on the manuscript.

CB, PH, DZ and MAJ were supported by a Wellcome Trust Senior Fellowship grant, number 087576, TEK and JW by the Leverhulme Trust, grant number F/00 212/AM, and PMD by a College of Medicine, Biological Sciences & Psychology studentship from the University of Leicester. The collection of the Frisian samples was supported by a grant to PdK from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. We thank Seamus O'Reilly, Michael Merrigan and Darren McGettigan of the Genealogical Society of Ireland for their support and participation in the recruitment of participants for the Irish DNA Atlas. AS was supported by the Finnish Foundations' Professor Pool (Paulo Foundation), JM by grant no. 175075 of the Ministry of Science of Serbia, AN by grant PRIN 2012JA4BTY_003, and CTS by Wellcome Trust grant no. 098051. MHDL is a postdoctoral fellow of the FWO-Vlaanderen (Research Foundation-Flanders). MHDL & AVG were supported by the KU Leuven BOF - Centre of Excellence Financing on 'Eco- and socio-evolutionary dynamics' (Project number PF/2010/07).

Figure legends

Figure 1: Distribution of sequenced regions on the MSY.

At the top is shown a schematic representation of the Y chromosome and the analysed sub-region, with the distribution of the ampliconic, X-transposed, X-degenerate and heterochromatic regions indicated (Skaletsky et al. 2003). The graph shows read depth in sequenced regions (blue) and density of discovered SNPs (red). Target coordinates for bait design (bottom) are according to GRCh37. Also shown are the locations of single-copy MSY genes (Skaletsky et al. 2003; Bellott et al. 2014), as triangles pointing in the direction of transcription. *TXLNGY* (Putative gamma-taxilin 2) replaces the former *CYorf15A* and *CYorf15B* (Skaletsky et al. 2003).

Figure 2: Venn diagram showing overlap of SNPs between next-generation sequencing studies of the MSY.

The total number of independent SNPs across all five studies (the current study plus Francalacci (Francalacci et al. 2013), Poznik (Poznik et al. 2013), Scozzari (Scozzari et al. 2014) and Wei (Wei et al. 2013a)) is 33,479.

Figure 3: Maximum parsimony tree of MSY SNP haplotypes.

(a) Major haplogroups are indicated by colours, and selected haplogroup-defining mutations are indicated on branches. Deep-rooting branches have been contracted for display. The coloured bar to the right indicates population group of origin: ASC: Asia, Central; ASE: Asia, East; BRI: British Isles; SCA: Scandinavia; ENW: Europe, North West; ESW: Europe, South West; ESC, Europe, South Central; ESE: Europe, South East; MNE: Middle & Near East; MEX: Mexico; AUS: Australia; AFP: Africa, food-producers; AHG: Africa, hunter-gatherers. Figure S1 gives tips labelled with individual sample names; (b) Simplified tree showing the true lengths for deep-rooting branches. Diagonal dashed lines indicate the positions of branch contractions in part (a).

Figure 4: Relationship between SNP- and STR-based TMRCA estimates.

SNP-based node estimates are plotted against STR-based estimates for (a) 21 STRs (b) 17 STRs and (c) 13 STRs, here using ASD with the 'ancestral haplotype' root specification. The black dashed line in each case indicates x=y. Underlying data and correlation coefficients are given in Tables S6 and S7.

Tables

Table 1: NGS studies of human Y chromosome diversity.

Study	Approach	Mb	Mean read depth	n	Sample choice	SNPs imputed?	SNPs found	Overlap with current study
1000GP pilot (1000 Genomes Project Consortium et al. 2010)	WGS	unclear	1.8 ×	77	4 HapMap populations	no (ML tree)	2870	635/13,261[a] (4.79%)
Wei (Wei et al. 2013a)	WGS	8.97	28.4 ×	36	various (Complete Genomics dataset, plus hg A male)	yes	5865 (+56 MNPs, 741 indels)	1776/13,261 (13.4%)
Poznik (Poznik et al. 2013)	WGS	9.9	median 3.1 × at var sites	69	9 populations (7 from HGDP)	yes	11,640	2420/13,261 (18.25%)
Francalacci (Francalacci et al. 2013)	WGS	8.97	2.16 ×	1208	Sardinian population	yes	11,763 (no singletons)	2229/13,261 (16.8%)
Scozzari (Scozzari et al. 2014)	SC	1.50	50 ×	68	phylogenetic	not stated	2386	665/13,261 (5.01%)
Current study	SC	3.7	51 ×	448	19 pops. + phylogenetic	no	13,261	**Novel: 8742/13,261 (65.9%)**

WGS: whole-genome sequence; SC: sequence capture; ML: maximum likelihood; HGDP: Human Genome Diversity Panel [a] based on available file containing 2788 variants in 75 individuals.

Table 2: TMRCA estimates for selected clades within the phylogeny.

Clade	N	TMRCA ± stdev/KYA	TMRCA range based on mutation rate CI/ KYA
root	448	125.8	50.3-419.5
B-M182	14	45.6	18.2-152.0
B2a-M150	2	16.6	6.7-55.5
B2b-M112	12	38.1	15.2-127.1
CR-P143	378	47.9	19.2-159.8
C-M216	9	39.4	15.8-131.5
DR-M168	427	48.7	19.5-162.4
D-M174	5	34.3	13.7-114.4
DE-M145	49	48.1	19.2-160.3
E-P29	44	37.9	15.2-126.4
E1b1a-M2	24	6.9	2.7-22.9
E1b1b-M215	17	17.7	7.1-58.9
FR-M213	369	35.2	14.1-117.2
HF5	7	31.6	12.7-105.5
G-M201	23	23.1	9.2-77.0
G2a-L31	20	16.4	6.6-54.8
H-M69	6	27.4	11.0-91.5
I-M170	76	20.6	8.2-68.6
I1-M253	46	3.5	1.4-11.5
I2-P215	30	17.1	6.8-57.0
J-M304	33	23.3	9.3-77.7
IJ-P123	109	31.0	12.4-103.4
J2-M172	28	21.1	8.4-70.3
J2a-M410	18	15.2	6.1-50.8
J2b-M102	10	11.3	4.5-37.8
LT	12	32.6	13.1-108.8
L-M11	5	14.2	5.7-47.3
T-M70	7	21.0	8.4-70.1
M/LT/NO/QR-M9	230	32.6	13.0-108.6
NO-M214	39	30.0	12.0-99.9
N-M231	20	13.4	5.4-44.8
N1c1-M178	15	4.6	1.8-15.2
O-P191	19	25.6	10.2-85.3
P-M45	179	24.2	9.7-80.6
Q-M242	5	22.6	9.0-75.4
R-M207	174	19.3	7.7-64.4
R1a-M198	27	6.2	2.5-20.8
R1b-L278	146	14.3	5.7-47.7
R1b-M269	145	4.9	2.0-16.3

References

1000 Genomes Project Consortium, RM Durbin, GR Abecasis, DL Altshuler, A Auton, LD Brooks, RA Gibbs, ME Hurles, GA McVean. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Balaresque, P, GR Bowden, EJ Parkin, et al. 2008a. Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis. Hum Mutat 29:1171-1180.

Balaresque, P, TE King, EJ Parkin, E Heyer, D Carvalho-Silva, T Kraaijenbrink, P de Knijff, C Tyler-Smith, MA Jobling. 2014. Gene conversion violates the stepwise mutation model for microsatellites in Y-chromosomal palindromic repeats. Hum Mutat 35:609-617.

Balaresque, P, EJ Parkin, L Roewer, et al. 2008b. Genomic complexity of the Y-STR DYS19: inversions, deletions and founder lineages carrying duplications. Int J Legal Med 123:15-23.

Ballantyne, KN, M Goedbloed, R Fang, et al. 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet 87:341-353.

Ballantyne, KN, V Keerl, A Wollstein, Y Choi, SB Zuniga, A Ralf, M Vermeulen, P de Knijff, M Kayser. 2012. A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Sci Int Genet 6:208-218.

Bandelt, H-J, P Forster, A Röhl. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48.

Batini, C, P Hallast, D Zadik, et al. submitted. Large-scale recent expansion of European patrilineages shown by population resequencing. Nature Comms.

Bellott, DW, JF Hughes, H Skaletsky, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature 508:494-499.

Berniell-Lee, G, F Calafell, E Bosch, E Heyer, L Sica, P Mouguiama-Daouda, L van der Veen, JM Hombert, L Quintana-Murci, D Comas. 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. Mol Biol Evol 26:1581-1589.

Bosch, E, ME Hurles, A Navarro, MA Jobling. 2004. Dynamics of a human interparalog gene conversion hotspot. Genome Res 14:835-844.

Busby, GB, F Brisighelli, P Sanchez-Diz, et al. 2012. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279:884-892.

Chang, X, K Wang. 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet 49:433-436.

Charchar, FJ, LD Bloomer, TA Barnes, et al. 2012. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. Lancet 379:915-922.

Drummond, AJ, A Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214.

Drummond, AJ, A Rambaut, B Shapiro, OG Pybus. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185-1192.

Elhaik, E, TV Tatarinova, AA Klyosov, D Graur. 2014. The 'extremely ancient' chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. Eur J Hum Genet 22:1111-1116.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author (Department of Genome Sciences, University of Washington, Seattle, WA).

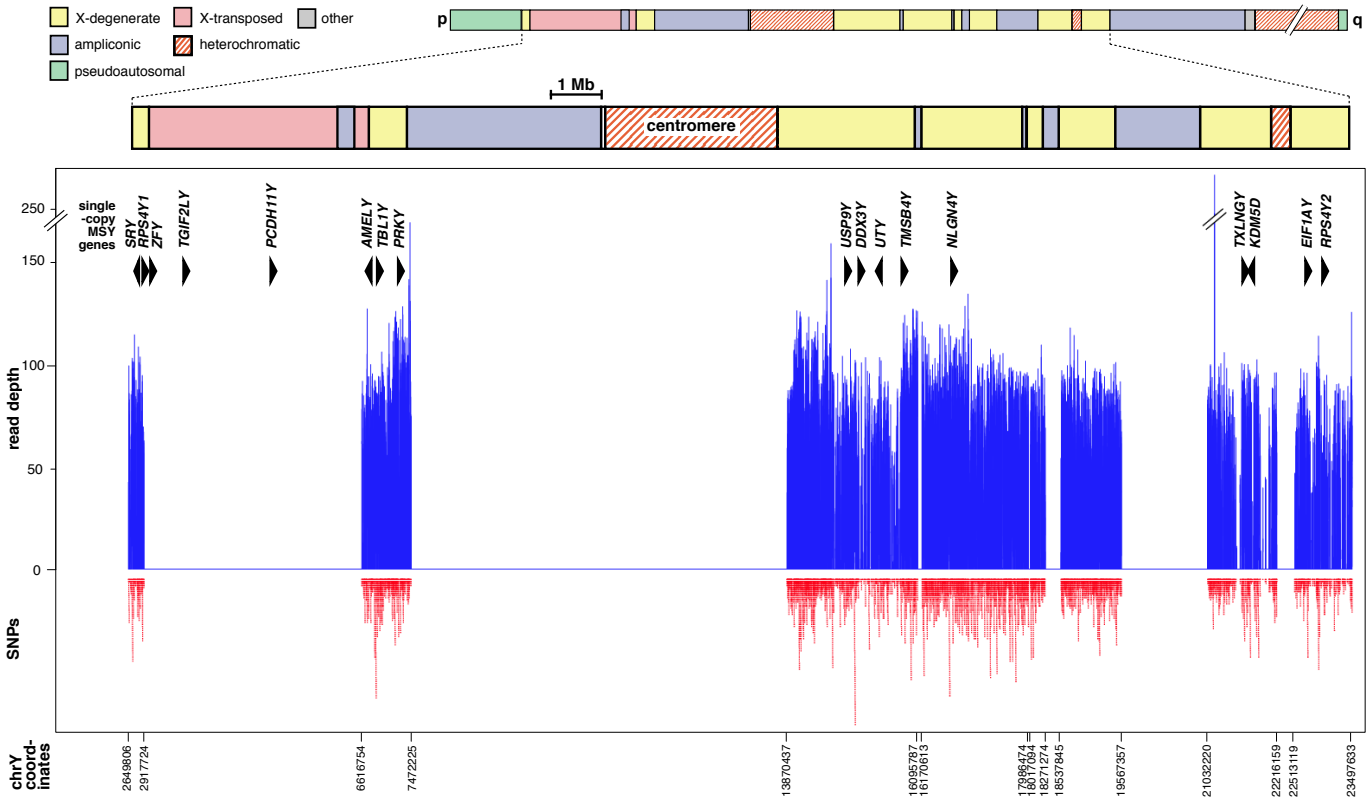
Forster, P, R Harding, A Torroni, H-J Bandelt. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935-945.

Francalacci, P, L Morelli, A Angius, et al. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341:565-569.

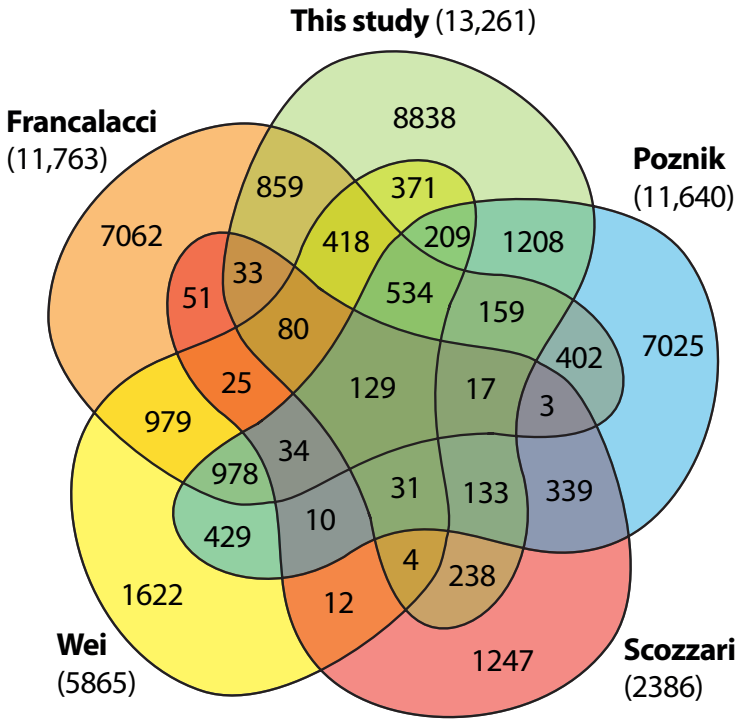
Goldstein, DB, AR Linares, LL Cavalli-Sforza, MW Feldman. 1995a. An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471.

Goldstein, DB, AR Linares, LL Cavalli-Sforza, MW Feldman. 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723-6727.
Hallast, P, P Balaresque, GR Bowden, SJ Ballereau, MA Jobling. 2013. Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. PLoS Genet 9:e1003666.
Hammer, MF. 1995. A recent common ancestry for human Y chromosomes. Nature 378:376-378.
Hammer, MF, DM Behar, TM Karafet, FL Mendez, B Hallmark, T Erez, LA Zhivotovsky, S Rosset, K Skorecki. 2009. Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. Hum Genet 126:707-717.
Jobling, MA, IC Lo, DJ Turner, et al. 2007. Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing *Amelogenin Y*. Hum Mol Genet 16:307-316.
Jobling, MA, C Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598-612.
Karafet, TM, FL Mendez, M Meilerman, PA Underhill, SL Zegura, MF Hammer. 2008. New binary polymorphisms reshape and increase resolution of the human Y-chromosomal haplogroup tree. Genome Res 18:830-838.
Karafet, TM, SL Zegura, O Posukh, et al. 1999. Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am J Hum Genet 64:817-831.
King, TE, EJ Parkin, G Swinfield, F Cruciani, R Scozzari, A Rosa, SK Lim, Y Xue, C Tyler-Smith, MA Jobling. 2007. Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. Eur J Hum Genet 15:288-293.
Kuroki, Y, A Toyoda, H Noguchi, et al. 2006. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nat Genet 38:158-167.
Larmuseau, MH, N Vanderheyden, A Van Geystelen, M van Oven, P de Knijff, R Decorte. 2014. Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Ann Hum Genet 78:92-103.
Mendez, FL, T Krahn, B Schrack, et al. 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 92:454-459.
Mendizabal, I, C Valente, A Gusmao, et al. 2011. Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. PLoS ONE 6:e15988.
Poznik, GD, BM Henn, MC Yee, et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562-565.
Purps, JS SiegertS Willuweit, et al. 2014. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Sci Int Genet 12C:12-23.
Rambaut, A. 2006-2012. Fig.Tree. Tree Figure Drawing Tool, version 1.4.0. Available at: http://tree.bio.ed.ac.uk/software/figtree/.
Repping, S, SK van Daalen, LG Brown, et al. 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet 38:463-467.
Rocca, RA, G Magoon, DF Reynolds, T Krahn, VO Tilroe, PM Op den Velde Boots, AJ Grierson. 2012. Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. PLoS ONE 7:e41634.
Rosser, ZH, P Balaresque, MA Jobling. 2009. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. Am J Hum Genet 85:130-134.
Rozen, S, JD Marszalek, RK Alagappan, H Skaletsky, DC Page. 2009. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. Am J Hum Genet 85:923-928.

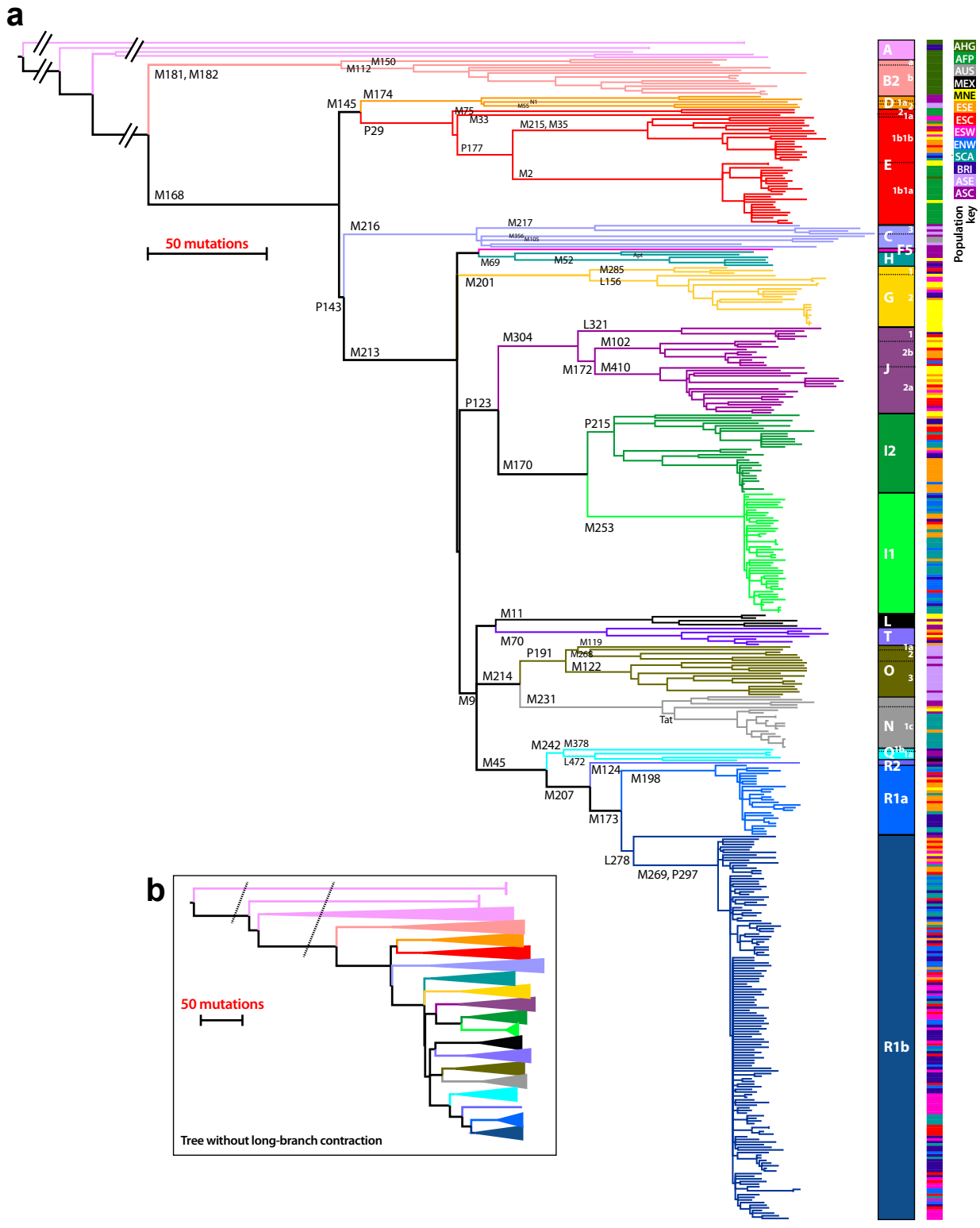
Rozen, S, H Skaletsky, JD Marszalek, PJ Minx, HS Cordum, RH Waterston, RK Wilson, DC Page. 2003. Abundant gene conversion between arms of massive palindromes in human and ape Y chromosomes. Nature 423:873-876.
Saillard, J, P Forster, N Lynnerup, H-J Bandelt, S Nørby. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718-726.
Scozzari, R, A Massaia, B Trombetta, G Bellusci, NM Myres, A Novelletto, F Cruciani. 2014. An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res 24:535-544.
Sezgin, E, JM Lind, S Shrestha, et al. 2009. Association of Y chromosome haplogroup I with HIV progression, and HAART outcome. Hum Genet 125:281-294.
Skaletsky, H, T Kuroda-Kawaguchi, PJ Minx, et al. 2003. The male-specific region of the human Y chromosome: a mosaic of discrete sequence classes. Nature 423:825-837.
Trombetta, B, F Cruciani, PA Underhill, D Sellitto, R Scozzari. 2010. Footprints of X-to-Y gene conversion in recent human evolution. Mol Biol Evol 27:714-725.
Trombetta, B, D Sellitto, R Scozzari, F Cruciani. 2014. Inter- and intraspecies phylogenetic analyses reveal extensive x-y gene conversion in the evolution of gametologous sequences of human sex chromosomes. Mol Biol Evol 31:2108-2123.
Underhill, PA, P Shen, AA Lin, et al. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-361.
Van Geystelen, A, R Decorte, MH Larmuseau. 2013a. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics 14:101.
Van Geystelen, A, R Decorte, MH Larmuseau. 2013b. Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs. Forensic Sci Int Genet 7:573-580.
van Oven, M, A Van Geystelen, M Kayser, R Decorte, MH Larmuseau. 2014. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat 35:187-191.
Wei, W, Q Ayub, Y Chen, S McCarthy, Y Hou, I Carbone, Y Xue, C Tyler-Smith. 2013a. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23:388-395.
Wei, W, Q Ayub, Y Xue, C Tyler-Smith. 2013b. A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci Int Genet 7:568-572.
Wilson Sayres, MA, KE Lohmueller, R Nielsen. 2014. Natural selection reduced diversity on human Y chromosomes. PLoS Genet 10:e1004064.
Xue, Y, Q Wang, Q Long, et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol 19:1453-1457.
Y Chromosome Consortium. 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12:339-348.
Zhivotovsky, LA, PA Underhill, C Cinnioglu, et al. 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74:50-61.



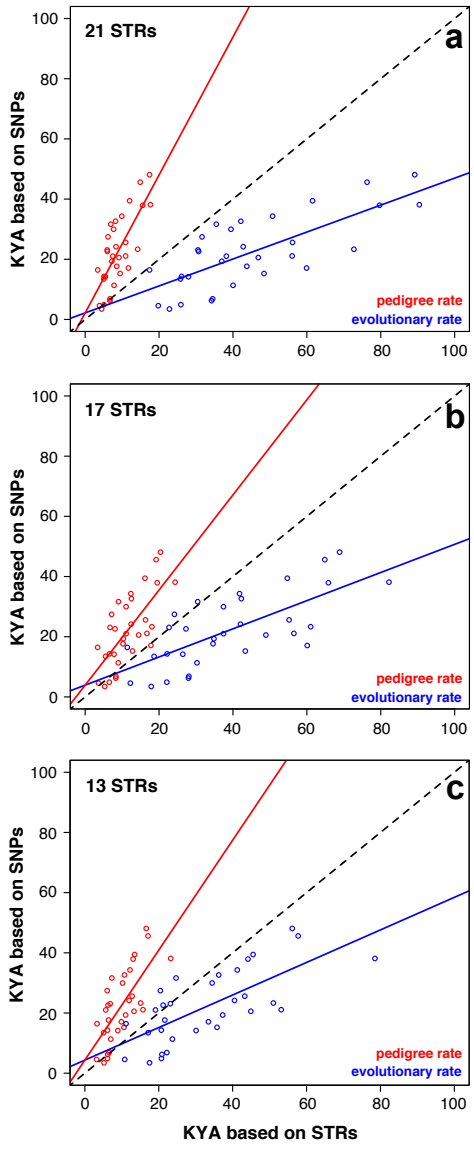
Hallast et al., Figure 1



Hallast et al., Figure 2



Hallast et al., Figure 3



Hallast et al., Figure 4